Better Backlinking

Semantische Kategorisierung von Websites

Stephan Sommer-Schulz
info@nerdbynature.net

Inhaltsübersicht

Ziel: Kategorien

Ziele:

- **Inhalte automatisch Kategorisieren**
- **Gezieltere / wirksamere Contentverlinkung**
- **Bessere Suchmaschinen**
- **Erkennung von Suchmaschinen-Spam**

Abgrenzung:

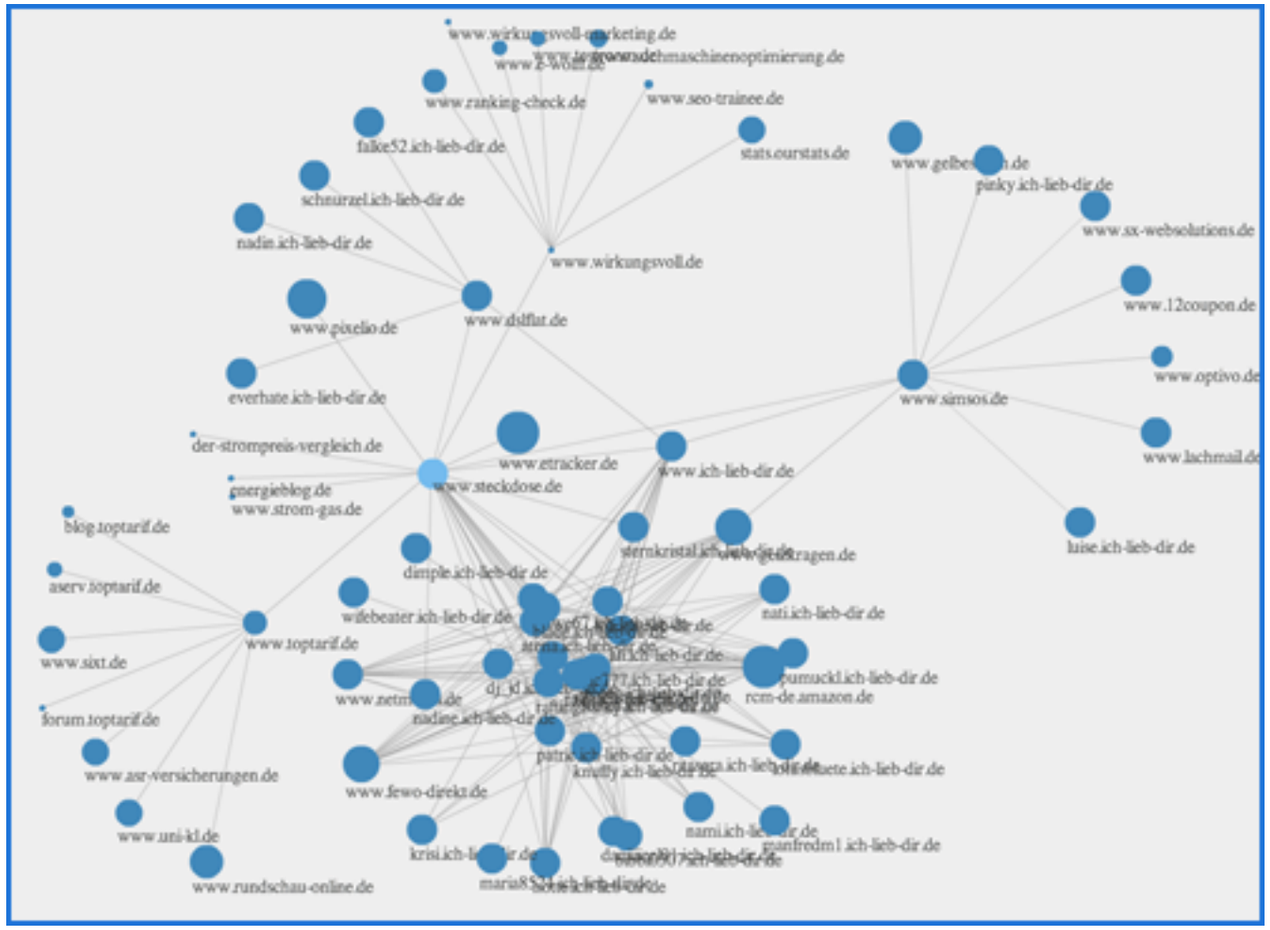
- Das Projekt hat nicht den Anspruch den Inhalt kompletter Webseiten automatisch zu „verstehen“.
- Semantik ist zwar ein Hype, kann aber dasselbe Schicksal wie KI in den 1980/90er ereilen.
- Watson: Rechenpower ohne Ende, beschränktes Einsatzscenario

Backlinks

Auswertung / Bewertung nach:

- Site-Herkunft
- Site-Rank
- Domain-Herkunft
- Geo-Verteilung
- IP-Netzverteilung
- Clusterbildung
- Keywords
- Linktypen

Neu: Kategorien / Semantik



Kategorien - Auswahl

Die Wahl der Kategorien kann / muss auf das jeweilige Umfeld angepasst werden

Beispiel:

- Bildung
- Computer und Internet
- Gastgewerbe
- Gesellschaft
- Gesundheit
- Kultur
- Nachrichten und Medien
- Online-Shops
- Regional
- Reise und Tourismus
- Spiele
- Sport
- Wissen
- Wissenschaft
- Zuhause

Semantische Verfahren

Manuelle Kategorisierung

Dieser Punkt erfreut die Herzen aller Praktikanten und Auszubildenden.

Ontologien

Ontologien sind Vokabulare mit denen bestehende Informationsangebote „manuell“ mit zusätzlichen Informationen detailliert bzw. erweitert werden können.

Verzeichnisse / Graphen

Viele Sites sind in großen Internet-Verzeichnissen wie dmoz oder Yahoo bereits kategorisiert.

Semantische Inhaltsanalyse

Die Inhalte einzelner Webseiten können semantisch analysiert werden. Somit ist eine Kategorisierung der Site möglich, ohne von „manuellen“ Fehlern oder Manipulationen abhängig zu sein.

Graphentheorie

Hypothese

Websites die miteinander Verlinkt sind gehören auch zu einem gewissen Maß thematisch zusammen.

Ableitung

Ein (Link-) Graph selbst mit mehreren Millionen Knoten kann thematisch erfasst werden, selbst dann, wenn nur für relativ wenige Knoten die Kategorien bekannt sind.

Voraussetzung

Es gibt definierte Gewichtungen für Knoten (Websites) und Kanten (Links) im Graphen

Webgraph - Datenbestand

Domains

Aktive DE-Domains (ca. 14.0 Mio. reg.)	2.829.193
Aktive CH-Domains (ca. 1.5 Mio. reg.)	367.792
Aktive AT-Domains (ca. 1.0 Mio. reg.)	213.316

Websites

Gefundene Websites	5.576.256
Korrekt erfasste Websites	5.107.121
Websites Fehler	433.567
Websites Timeout	7.158

Webpages

Einzelne Webpages erfasst	53.568.238

Links

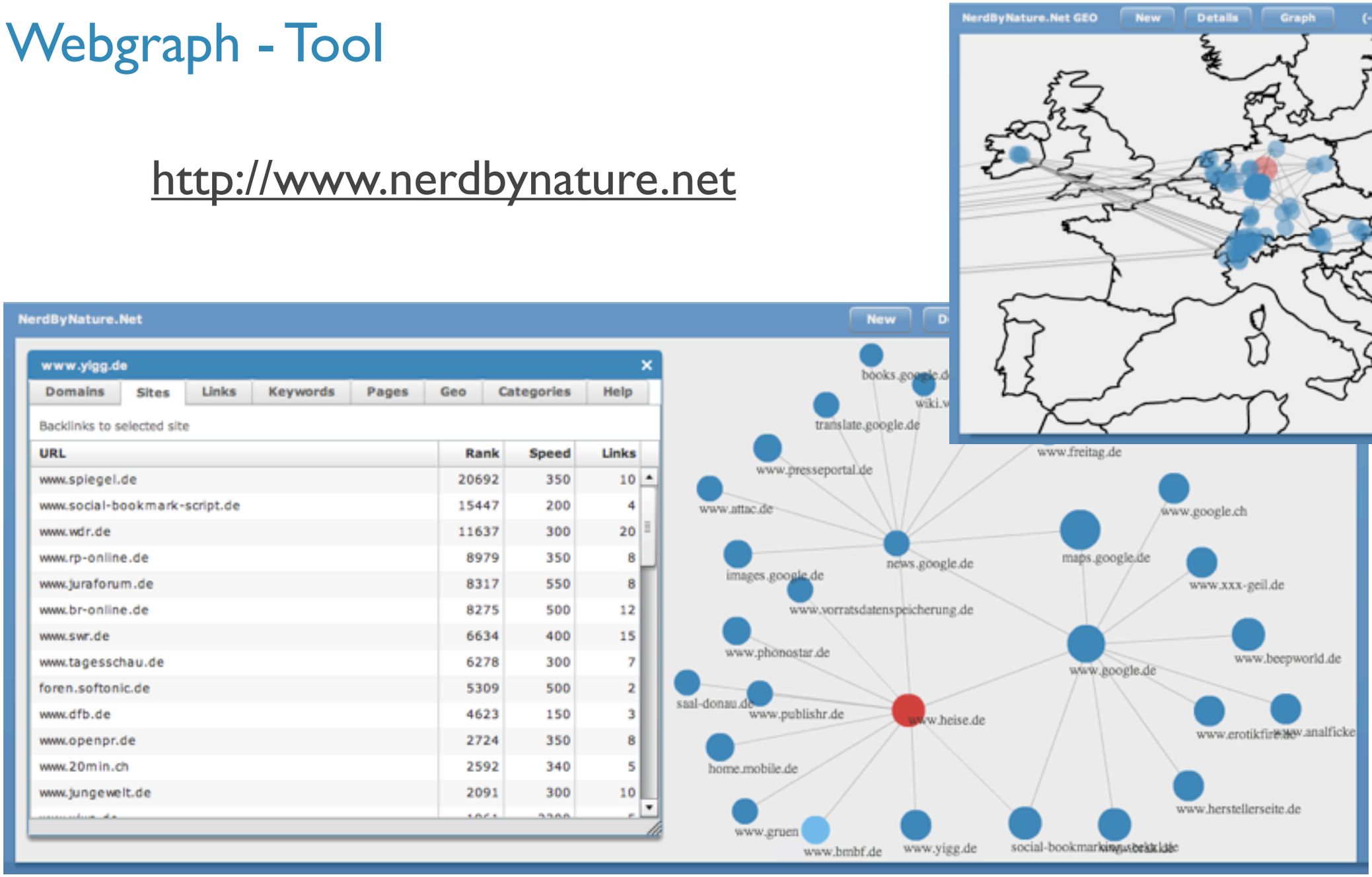
Website zu Website Links	239.190.607
Webpage Deeplinks	445.389.631

Kategorien - Datenbestand

Kategorien	Anzahl Websites
• Bildung	7.588
• Computer und Internet	6.418
• Gastgewerbe	4.101
• Gesellschaft	6.010
• Gesundheit	21.810
• Kultur	27.852
• Nachrichten und Medien	3.404
• Online-Shops	14.012
• Regional	4.876
• Reise und Tourismus	1.944
• Spiele	4.232
• Sport	27.476
• Verkehr	876
• Verwaltung	127
• Wirtschaft	57.453
• Wissen und Wissenschaft	14.177
• Zuhause	1.814
Summe Sites:	**204.170**

Webgraph - Tool

http://www.nerdbynature.net



Forschung (Ausblick)

Mit Unterstützung des europäischen Fonds für regionale Entwicklung (EFRE), entwickelt die W3 Solutions GmbH eine **Meme-Suche** der nächsten Generation. Erste Teilergebnisse der bisherigen Forschungstätigkeit sind:

1. Word Connections Finder

   Finde Wortverbindungen zum Suchwort

2. Keyword Finder

   Texte autom. nach Stichwörtern durchsuchen

3. Neighbour Finder

   Relevante Wortnachbarn und Ketten finden

4. Similar Graph Finder

   Ähnliche Worte über semantische Graphen finden

5. Sentance Tagger

   Grammatikalische Bestandteile von Texten erkennen

6. Word Suggestions

   Wortvorschläge generieren (in Vorbereitung)

7. Search Cloud, Trend Finder

   Aktuelle Suchtrends erkennen

Kontakt

NerdByNature.Net

Stephan Sommer-Schulz

Alt-Wittenau 37H

13437 Berlin

Tel: +49 30 3450595-0

Fax: +49 30 3450595-11

Mail: info@nerdbynature.net

Web: http://www.nerdbynature.net